- SVM
  - pre-requirement
    - vector calculus
    - vector projection
    - distance between points and lines
    - Lagrange Optimization
      - multiplier
      - dual representation
  - SVM
    - linearly seperable
    - not linearly seperable

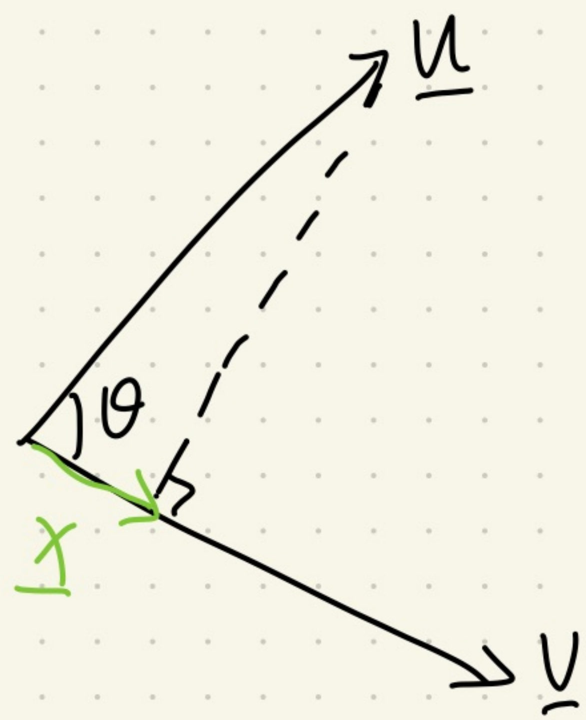
1. vector calculus

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0 \quad \text{or} \quad = \underline{x}^T \underline{w} + w_0 \Rightarrow \nabla_{\underline{x}}\, g(\underline{x}) = \underline{w} \quad \text{because} \begin{cases} \nabla_{\underline{x}}(\underline{w}^T \underline{x}) = \underline{w} & ① \\ \nabla_{\underline{x}}(\underline{x}^T \underline{w}) = \underline{w} & ② \end{cases}$$

prf: ① $\underline{w}^T \underline{x} = \sum_{i=1}^{n} w_i x_i \Rightarrow \nabla_{\underline{x}}(\underline{w}^T \underline{x}) = \nabla_{\underline{x}}\left(\sum_{i=1}^{n} w_i x_i\right) = \begin{bmatrix} \frac{\partial\left(\sum_{i=1}^{n} w_i x_i\right)}{\partial x_1} \\ \vdots \\ \frac{\partial\left(\sum_{i=1}^{n} w_i x_i\right)}{\partial x_n} \end{bmatrix}_{n\times 1} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \underline{w}$

② $\underline{x}^T \underline{w} = \sum_{i=1}^{n} w_i x_i$ same to above

2. vector projection



$$\underline{x} = \{\text{magnitude}\} \cdot \{\text{direction}\}$$

$$= \{\|\underline{u}\| \cdot \cos\theta\} \cdot \left\{\frac{\underline{v}}{\|\underline{v}\|}\right\}$$

unit vector

$$= \|\underline{u}\| \cdot \frac{\underline{u}^T \cdot \underline{v}}{\|\underline{u}\| \cdot \|\underline{v}\|} \cdot \frac{\underline{v}}{\|\underline{v}\|}$$

$$= \underline{u}^T \cdot \frac{\underline{v}}{\|\underline{v}\|^2} \cdot \underline{v} = \frac{\underline{u}^T \cdot \underline{v}}{\underline{v}^T \underline{v}} \cdot \underline{v}$$

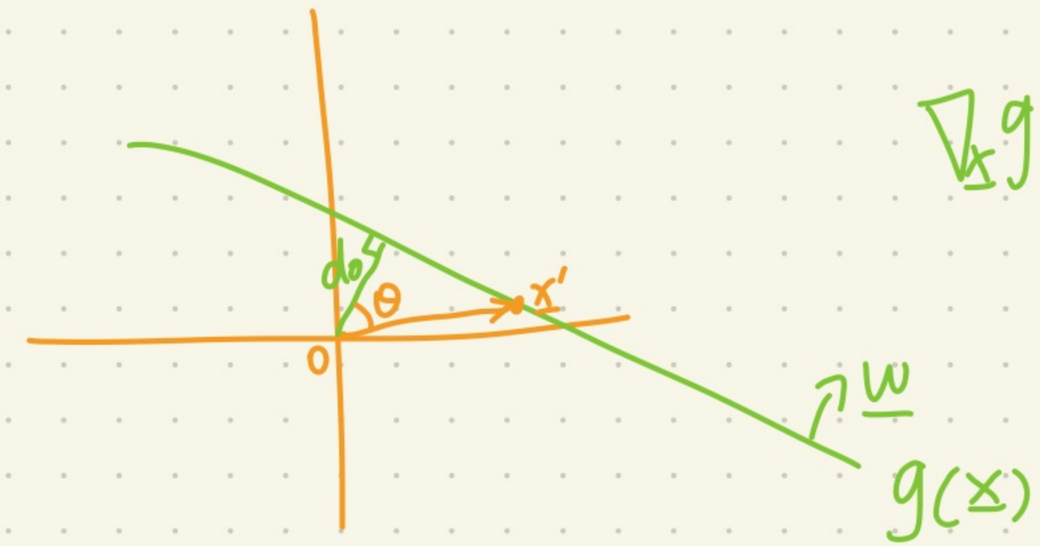
3. Distance

$g(\underline{x}) = \underline{w}^T\underline{x} + w_0$

$\nabla_{\underline{x}} g(\underline{x}) \perp g(\underline{x}) \Rightarrow \underline{w} \perp g(\underline{x})$

① origin to $g(\underline{x})=0$

(not general)



find any point $\underline{x}'$ on $g(\underline{x})=0$, so $\begin{cases} g(\underline{x}')=0 \\ \underline{w}^T\underline{x}'+w_0=0 \end{cases}$

$$d_0 = \|\underline{x}'\| \cdot \cos\theta = \|\underline{x}'\| \cdot \frac{\underline{w}^T \cdot \underline{x}'}{\|\underline{w}\| \cdot \|\underline{x}'\|} = \frac{\underline{w}^T\underline{x}'}{\|\underline{w}\|} = \boxed{\frac{-w_0}{\|\underline{w}\|}} \quad \therefore \text{QED}$$

$\frac{-w_0}{\|w\|}$

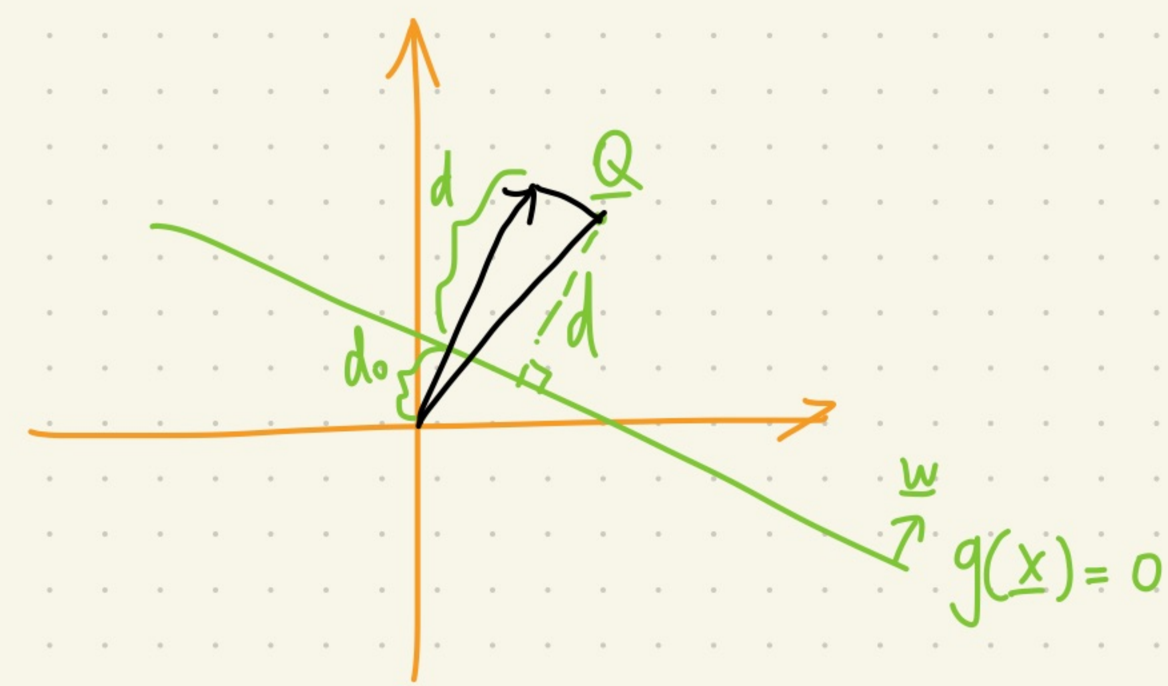
↳ here we actually get the absolute value of d_0, so we have the minus.

② any point to $g(x)=0$ (general formula)

$$d = \underbrace{d\{\text{Origin to } \underline{Q} \text{ in } \underline{w} \text{ direction}\}}_{\|\underline{Q}\| \text{ projects to } \underline{w}} - d\{\text{Origin to } g(\underline{x})=0\}$$

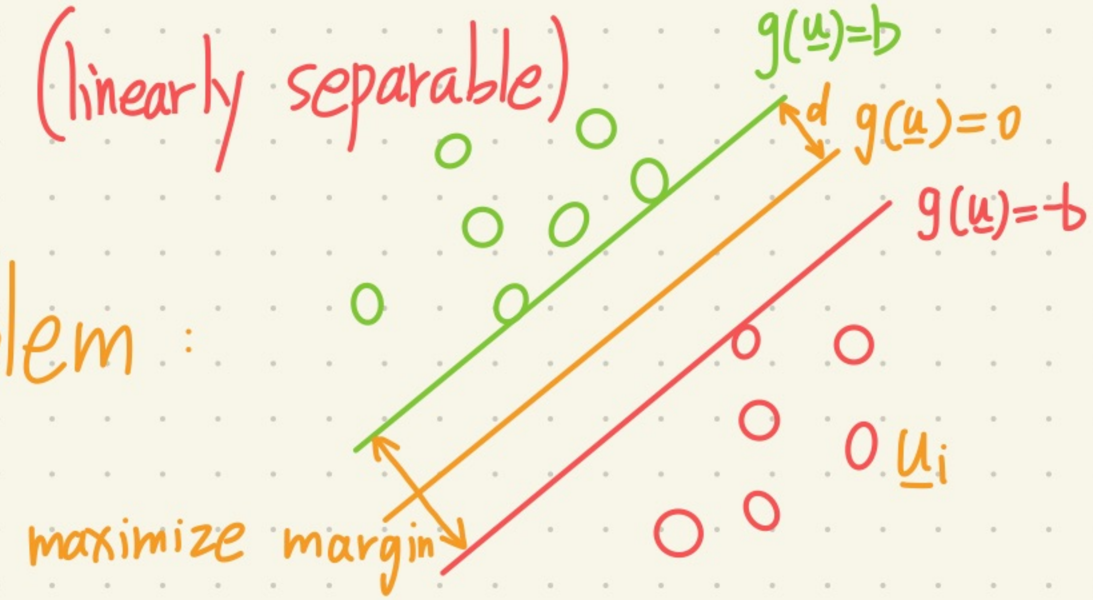
$$= \|\underline{Q}\| \cdot \frac{\underline{w}^T \cdot \underline{Q}}{\|\underline{w}\| \cdot \|\underline{Q}\|} - \frac{-w_0}{\|\underline{w}\|}$$

$$= \frac{\underline{w}^T \cdot \underline{Q} + w_0}{\|\underline{w}\|} = \boxed{\frac{g(\underline{Q})}{\|\underline{w}\|}} \quad \therefore \text{QED}$$



4. SVM (linearly separable)

(1) problem:



Basically, we need

① $\begin{cases} z_i d\{\underline{u}_i, g(\underline{x})=0\} = \dfrac{z_i(\underline{w}^T\underline{u}_i + w_0)}{\|\underline{w}\|} \geq d > 0, \ \forall i \\ \text{maximize } d \text{ as possible} \end{cases}$

We have $d = \dfrac{g(\underline{u}')}{\|\underline{w}\|} = \dfrac{b}{\|\underline{w}\|}$, then $\begin{cases} \text{maximize } b \\ \text{minimize } \|\underline{w}\| \end{cases}$

on $g(\underline{u}) = b$

① becomes $z_i(\underline{w}^T\underline{u}_i + w_0) \geq b > 0, \ \forall i$ with $\begin{cases} \text{minimizing } \|\underline{w}\| \\ \text{maxizing } b \end{cases}$

Now, consider b as a constant, we only minimize $\|\underline{w}\|$.

Therefore,

$$
② \begin{cases} \text{minimize } J(\underline{w}) = \frac{1}{2}\|\underline{w}\|^2 & \text{(better for calculation)} \\ \text{constraints: } z_i(\underline{w}^T\underline{u}_i + w_0) - b \geq 0, & \forall i \end{cases}
$$

(2) Lagrange Optimization (use $\underline{\lambda}$ for lagrange multipliers)

$$
\boxed{L(\underline{w}, w_0, \underline{\lambda}) = \frac{1}{2}\|\underline{w}\|^2 - \sum_{i=1}^{N} \lambda_i [z_i(\underline{w}^T\underline{u}_i + w_0) - b]} \quad \text{Primal form}
$$

$$
\boxed{\text{KKT conditions} \begin{cases} \lambda_i \geq 0 & \forall i \\ \lambda_i [z_i(\underline{w}^T\underline{u}_i + w_0) - b] = 0 & \forall i \\ z_i(\underline{w}^T\underline{u}_i + w_0) - b \geq 0 & \forall i \end{cases}}
$$

Now, we need to calculate $\min_{\underline{w}, w_0} \max_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$

The intuition about $\max_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$ is punishing outliers

$$\begin{cases} \text{if any } z_i(\underline{w}^T\underline{u}_i + w_0) - b < 0, \text{ we want } L(\underline{w}, w_0, \underline{\lambda}) \text{ be bigger as punishment,} \\ \qquad \text{so } \lambda_i \text{ need to} \to \infty, L \to \infty \\ \text{if all } z_i(\underline{w}^T\underline{u}_i + w_0) - b \geq 0, \text{ according to KKT, } \lambda_i = 0 \text{ for } \forall i, L = \frac{1}{2}\|\underline{w}\|^2 \\ \qquad \text{(no punishment)} \end{cases}$$

(3) Dual Representation [General]

Change calculation from

$$\min_{\underline{w}, w_0} \max_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$$

to

$$\max_{\underline{\lambda}} \min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda})$$

try to not change

If we don't do that

$$L(\underline{w}, w_0, \underline{\lambda}) = \frac{1}{2}\|\underline{w}\|^2 - \sum_{i=1}^{N} \lambda_i \left[\gamma_i (\underline{w}^T \underline{u}_i + w_0) - b \right]$$

$$\begin{cases} \nabla_{\lambda_i} L(\underline{w}, w_0, \underline{\lambda}) = -\gamma_i (\underline{w}^T \underline{u}_i + w_0) + b \\ \nabla_{\lambda_i}^2 L(\underline{w}, w_0, \underline{\lambda}) = 0 \end{cases}$$

We can see there is no way to get λ_i

Hence, we need to calculate

$$\max_{\underline{\lambda}} \min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda})$$

WHY this works?

① $\max\limits_{\underline{\lambda}} \min\limits_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda}) \le \min\limits_{\underline{w}, w_0} \max\limits_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$

prf: obviously,

$$\min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda}) \le L(\underline{w}, w_0, \underline{\lambda}) \le \max_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$$

So

$$\min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda}) \le \max_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$$

$$\max_{\underline{\lambda}} \min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda}) \le \min_{\underline{w}, w_0} \max_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$$

② make $\max\limits_{\underline{\lambda}} \min\limits_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda}) = \min\limits_{\underline{w}, w_0} \max\limits_{\underline{\lambda}} L(\underline{w}, w_0, \underline{\lambda})$

We need new KKT conditions:

$$\text{KKT conditions} \begin{cases} \left.\begin{array}{ll} \lambda_i \geq 0 & \forall i \\ \lambda_i \left[z_i (\underline{w}^T \underline{u}_i + w_0) - b \right] = 0 & \forall i \\ z_i (\underline{w}^T \underline{u}_i + w_0) - b \geq 0 & \forall i \end{array} \right\} \text{original} \\ \left.\begin{array}{l} \nabla_{\underline{w}} L(\underline{w}, w_0, \underline{\lambda}) = 0 \\ \nabla_{w_0} L(\underline{w}, w_0, \underline{\lambda}) = 0 \end{array} \right\} \text{new} \end{cases}$$

Now, we can solve the problem by calculating

$$\max_{\underline{\lambda}} \min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda})$$

(4) Dual Representation [Specifically]

① calculate $\min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda})$

According to new KKT $L(\underline{w}, w_0, \underline{\lambda}) = \frac{1}{2}\|\underline{w}\|^2 - \sum_{i=1}^{N} \lambda_i [z_i(\underline{w}^T \underline{u}_i + w_0) - b]$

New 1: $\nabla_{\underline{w}} L(\underline{w}, w_0, \underline{\lambda}) = 0 \Rightarrow \nabla_{\underline{w}} \left(\frac{1}{2} \underline{w}^T \underline{w} - \sum_{i=1}^{N} \lambda_i [z_i(\underline{w}^T \underline{u}_i + w_0) - b] \right)$

$$= \frac{1}{2}\underline{w} + \frac{1}{2}\underline{w} - \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i = 0$$

$$\therefore \underline{w}^* = \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i \quad (i)$$

(To ensure this is minimizing, not maximizing, check $\nabla^2_{\underline{w}} L > 0$)

$\nabla^2_{\underline{w}} = \nabla_{\underline{w}} \left(\underline{w} - \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i \right) = 1 > 0$ $\quad \therefore$ Yes, it's minimum.

New 2: $\nabla_{w_0} L(\underline{w}, w_0, \underline{\lambda}) = 0 \Rightarrow \nabla_{w_0}\left(\frac{1}{2}\underline{w}^T\underline{w} - \sum_{i=1}^{N} \lambda_i [z_i(\underline{w}^T\underline{u}_i + w_0) - b]\right)$

$$= -\sum_{i=1}^{N} \lambda_i z_i = 0$$

$$\therefore \sum_{i=1}^{N} \lambda_i z_i = 0 \quad \text{(ii)}$$

$\nabla_{w_0}^2 = \nabla_{w_0}\left(-\sum_{i=1}^{N} \lambda_i z_i\right) = 0$ $\quad\therefore$ It's not maximizing

Now we bring the results (i) & (ii) back to $L(\underline{w}, w_0, \underline{\lambda})$, we get the dual representation of L:

$$L_D(\underline{\lambda}) = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i z_i \underline{u}_i\right)^T \cdot \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i + \sum_{i=1}^{N} \lambda_i b - \sum_{i=1}^{N} \lambda_i z_i \left[\left(\sum_{i=1}^{N} \lambda_i z_i \underline{u}_i\right)^T \underline{u}_i + w_0\right]$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T\underline{u}_j + \sum_{i=1}^{N}\lambda_i b - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underbrace{\underline{u}_i^T\underline{u}_j} - \underbrace{\sum_{i=1}^{N}\lambda_i z_i}_{=0} \cdot w_0$$

Note: $\underline{u}_i^T\underline{u}_j = \underline{u}_j^T\underline{u}_i$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T\underline{u}_j + \sum_{i=1}^{N}\lambda_i b$$

dual representation

Now, we finished $\min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda})$ and changed it to $L_D(\underline{\lambda})$

$$\boxed{\begin{gathered}\max_{\underline{\lambda}} \min_{\underline{w}, w_0} L(\underline{w}, w_0, \underline{\lambda}) \\ \Downarrow \\ \max_{\underline{\lambda}} L_D(\underline{\lambda})\end{gathered}}$$

② calculate $\max\limits_{\underline{\lambda}} L_D(\underline{\lambda})$ [Algebraic Method]

We have $L_D(\underline{\lambda}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T \underline{u}_j + \sum_{i=1}^{N}\lambda_i b$

Add one more term $\mu\sum_{i=1}^{N}\lambda_i z_i$ to it

(in order to make $\underline{\lambda}$ satisfy KKT : $\sum_{i=1}^{N}\lambda_i z_i = 0$)

new: $L_D(\underline{\lambda},\mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T \underline{u}_j + \sum_{i=1}^{N}\lambda_i b + \underbrace{\mu}_{\text{new lagrange multiplier}}\sum_{i=1}^{N}\lambda_i z_i$

$$\nabla_{\lambda_k} L_D(\underline{\lambda}) = -\frac{1}{2}\sum_{\substack{j=1\\(j\neq k)}}^{N}\lambda_j z_k z_j \underline{u}_k^T \underline{u}_j - \frac{1}{2}\sum_{\substack{i=1\\(i\neq k)}}^{N}\lambda_i z_i z_k \underline{u}_i^T \underline{u}_k - \lambda_k z_k z_k \underline{u}_k^T \underline{u}_k + b + \mu z_k$$

$$= -\sum_{\substack{i=1 \\ (i\neq k)}}^{N} \lambda_i z_i z_k \underline{u}_i^T \underline{u}_k - \lambda_k z_k z_k \underline{u}_k^T \underline{u}_k + b + \mu z_k$$

$$= -\sum_{i=1}^{N} \lambda_i z_i z_k \underline{u}_i^T \underline{u}_k + b + \mu z_k = 0$$

Now, $\sum_{i=1}^{N} \lambda_i z_i z_k \underline{u}_i^T \underline{u}_k = \mu z_k + b$ for λ_k in $\underline{\lambda}$

$$k=1 \quad z_1 \underline{u}_1^T z_1 \underline{u}_1 \cdot \lambda_1 + z_2 \underline{u}_2^T z_1 \underline{u}_1 \cdot \lambda_2 + \cdots + z_N \underline{u}_N^T z_1 \underline{u}_1 \cdot \lambda_N = \mu z_1 + b$$

$$k=2 \quad z_1 \underline{u}_1^T \cdot z_2 \underline{u}_2 \cdot \lambda_1 + z_2 \underline{u}_2^T z_2 \underline{u}_2 \cdot \lambda_2 + \cdots + z_N \underline{u}_N^T z_2 \underline{u}_2 \cdot \lambda_N = \mu z_2 + b$$

$$\Rightarrow \quad \vdots \qquad \qquad \vdots$$

$$k=N \quad z_1 \underline{u}_1^T \cdot z_N \underline{u}_N \cdot \lambda_1 + z_2 \underline{u}_2^T z_N \underline{u}_N \cdot \lambda_2 + \cdots + z_N \underline{u}_N^T z_N \underline{u}_N \cdot \lambda_N = \mu z_N + b$$

$$z_1 \cdot \lambda_1 + z_2 \cdot \lambda_2 + \cdots + z_N \cdot \lambda_N = 0$$

$$\Rightarrow \begin{bmatrix} z_1 \underline{u}_1^T z_1 \underline{u}_1 & z_2 \underline{u}_2^T z_1 \underline{u}_1 & \cdots & z_N \underline{u}_N^T z_1 \underline{u}_1 \\ z_1 \underline{u}_1^T z_2 \underline{u}_2 & z_2 \underline{u}_2^T z_2 \underline{u}_2 & \cdots & z_N \underline{u}_N^T z_2 \underline{u}_2 \\ \vdots & \vdots & & \vdots \\ z_1 \underline{u}_1^T z_N \underline{u}_N & z_2 \underline{u}_2^T z_N \underline{u}_N & \cdots & z_N \underline{u}_N^T z_N \underline{u}_N \\ z_1 & z_2 & & z_N \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{bmatrix} = \mu \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \\ 0 \end{bmatrix} + \begin{bmatrix} b \\ b \\ \vdots \\ b \\ 0 \end{bmatrix}$$

$$(N+1) \times N \qquad N \times 1 \qquad (N+1) \times 1 \qquad (N+1) \times 1$$

Add μ to $\underline{\lambda}$

$$\Rightarrow \begin{bmatrix} z_1 \underline{u}_1^T z_1 \underline{u}_1 & z_2 \underline{u}_2^T z_1 \underline{u}_1 & \cdots & z_N \underline{u}_N^T z_1 \underline{u}_1 & -z_1 \\ z_1 \underline{u}_1^T z_2 \underline{u}_2 & z_2 \underline{u}_2^T z_2 \underline{u}_2 & \cdots & z_N \underline{u}_N^T z_2 \underline{u}_2 & -z_2 \\ \vdots & \vdots & & \vdots & \vdots \\ z_1 \underline{u}_1^T z_N \underline{u}_N & z_2 \underline{u}_2^T z_N \underline{u}_N & \cdots & z_N \underline{u}_N^T z_N \underline{u}_N & -z_N \\ z_1 & z_2 & & z_N & 0 \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \mu \end{bmatrix} = \begin{bmatrix} b \\ b \\ \vdots \\ b \\ 0 \end{bmatrix}$$

$$(N+1) \times (N+1) \qquad (N+1) \times 1 \qquad (N+1) \times 1$$

Now, we let $\underline{\underline{A}}_{(N+1)\times(N+1)}$ $\qquad$ $\underline{\rho}_{(N+1)\times 1}$ $\qquad$ $\underline{b}_{(N+1)\times 1}$

So we get $\underline{\underline{A}} \cdot \underline{\rho} = \underline{b}$

Solve $\underline{\rho} = \underline{\underline{A}}^{-1} \underline{b}$, Then we get $\underline{\lambda}^*$ & u^*

Solve $\underline{w}^* = \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i$ with $\underline{\lambda}^*$

Find one point $\underline{u}'$ on the margin $(g(\underline{u}') = \pm b)$

Use $\lambda_i [z_i (\underline{w}^T \underline{u}_i + w_0) - b] = 0 \implies z' (\underline{w}^{*T} \underline{u}' + w_0^*) - b = 0$

Solve $w_0^* = \frac{b}{z'} - \underline{w}^{*T} \cdot \underline{u}'$ with $b, z', \underline{w}^*$ & $\underline{u}'$ $\qquad$ $\therefore$ Done

Comment:

i) We use the linear algebra approach to solve the problem, which has no way of enforcing the requirement $\lambda_i \geq 0 \ \forall i$.

ii) If any $\lambda_k < 0$, we can set $\lambda_k = 0$ in $\underline{\ell}$, then recalculate $\underline{\ell} = \underline{\underline{A}}^{-1}\underline{b}$ ($\underline{\underline{A}}$ becomes $N \times N$ from $N+1 \times N+1$, $\underline{b} \to N \times 1$, $\underline{\ell} \to N \times 1$ if set one λ)

iii) Only worked for num of points ≤ 3.

③ calculate $\max_{\underline{\lambda}} L_D(\underline{\lambda})$ [SMO Method]

$$\boxed{L_D(\underline{\lambda}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T \underline{u}_j + \sum_{i=1}^{N}\lambda_i b}$$

i) each time choose one pair $\langle \lambda_m, \lambda_n \rangle$

ii) $$L_D(\lambda_m,\lambda_n) = (\lambda_m+\lambda_n)b + \sum_{\substack{i=1\\(i\neq m,n)}}^{N}\lambda_i b - \underbrace{\frac{1}{2}\lambda_m\lambda_n z_m z_n \underline{u}_m^T\underline{u}_n}_{i=m,\, j=n} - \underbrace{\frac{1}{2}\lambda_m\lambda_m z_m z_m \underline{u}_m^T\underline{u}_m}_{i=m,\, j=m}$$

$$- \underbrace{\frac{1}{2}\lambda_n\lambda_m z_n z_m \underline{u}_n^T\underline{u}_m}_{i=n,\, j=m} - \underbrace{\frac{1}{2}\lambda_n\lambda_n z_n z_n \underline{u}_n^T\underline{u}_n}_{i=n,\, j=n} - \frac{1}{2}\lambda_m z_m \sum_{\substack{j=1\\(j\neq m,n)}}^{N}\lambda_j z_j \underline{u}_m^T\underline{u}_j \;(i=m)$$

$$-\frac{1}{2}\lambda_n z_n \sum_{\substack{j=1\\(j\neq m,n)}}^{N}\lambda_j z_j \underline{u}_n^T\underline{u}_j \;(i=n) - \frac{1}{2}\lambda_m z_m \sum_{\substack{i=1\\(i\neq m,n)}}^{N}\lambda_i z_i \underline{u}_i^T\underline{u}_m \;(j=m) - \frac{1}{2}\lambda_n z_n \sum_{\substack{i=1\\(i\neq m,n)}}^{N}\lambda_i z_i \underline{u}_i^T\underline{u}_n \;(j=n)$$

$$= (\lambda_m+\lambda_n)b - \frac{1}{2}\lambda_m\lambda_m z_m z_m \underline{u}_m^T \underline{u}_m - \frac{1}{2}\lambda_n\lambda_n z_n z_n \underline{u}_n^T \underline{u}_n - \lambda_m\lambda_n z_m z_n \underline{u}_m^T \underline{u}_n$$

$$- \lambda_m z_m \sum_{\substack{i=1 \\ (i\neq m,n)}}^{N} \lambda_i z_i \underline{u}_i^T \underline{u}_m - \lambda_n z_n \sum_{\substack{i=1 \\ (i\neq m,n)}}^{N} \lambda_i z_i \underline{u}_i^T \underline{u}_n + \underbrace{\sum_{i=1}^{N} \lambda_i b}_{=C}$$

let $K_{pq} = \underline{u}_p^T \underline{u}_q$

$$\boxed{L_D(\lambda_m,\lambda_n) = (\lambda_m+\lambda_n)b - \frac{1}{2}\lambda_m^2 z_m^2 K_{mm} - \frac{1}{2}\lambda_n^2 z_n^2 K_{nn} - \lambda_m\lambda_n z_m z_n K_{mn} - \lambda_m z_m \sum_{\substack{i=1 \\ (i\neq m,n)}}^{N} \lambda_i z_i K_{im} - \lambda_n z_n \sum_{\substack{i=1 \\ (i\neq m,n)}}^{N} \lambda_i z_i K_{in} + C}$$

iii) get equations pertaining to λ_m, λ_n by $\underbrace{\sum_{i=1}^{N} \lambda_i z_i = 0}_{\text{KKT condition}}$

$$\lambda_m z_m + \lambda_n z_n = \underbrace{-\sum_{\substack{i=1 \\ (i\neq m,n)}}^{N} \lambda_i z_i}_{=\zeta}$$

$$\lambda_m z_m = \varsigma - \lambda_n z_n$$

$$\lambda_m = \varsigma \cdot z_m - \lambda_n z_n z_m \qquad \text{because } z_k \cdot z_k = 1$$

Let $V_m = \sum_{\substack{i=1 \\ (i \neq m,n)}}^{N} \lambda_i z_i K_{im}$ $\qquad V_n = \sum_{\substack{i=1 \\ (i \neq m,n)}}^{N} \lambda_i z_i K_{in}$

$$L_D(\lambda_m, \lambda_n) = (\lambda_m + \lambda_n) b - \frac{1}{2}\lambda_m^2 K_{mm} - \frac{1}{2}\lambda_n^2 K_{nn} - \lambda_m \lambda_n z_m z_n K_{mn} - \lambda_m z_m V_m - \lambda_n z_n V_n + C$$

$$\Longrightarrow \lambda_m = \varsigma \cdot z_m - \lambda_n z_n z_m$$

$$\begin{aligned}\therefore L_D(\lambda_n) = {} & (\varsigma \cdot z_m - \lambda_n z_n z_m + \lambda_n) b - \frac{1}{2}(\varsigma z_m - \lambda_n z_n z_m)^2 K_{mm} - \frac{1}{2}\lambda_n^2 K_{nn} \\ & - (\varsigma z_m - \lambda_n z_n z_m)\lambda_n z_m z_n K_{mn} - (\varsigma z_m - \lambda_n z_n z_m) z_m V_m - \lambda_n z_n V_n + C\end{aligned}$$

$$\boxed{\begin{aligned} = {} & -\frac{1}{2}(\varsigma - \lambda_n z_n)^2 K_{mm} - \frac{1}{2}\lambda_n^2 K_{nn} - (\varsigma - \lambda_n z_n)\lambda_n z_n K_{mn} - (\varsigma - \lambda_n z_n) V_m - \lambda_n z_n V_n \\ & + (\varsigma z_n - \lambda_n z_n z_m + \lambda_n) b + C\end{aligned}}$$

iv) maximize $L_D(\lambda_n)$

$$\frac{\partial L_D(\lambda_n)}{\partial \lambda_n} = -(\xi-\lambda_n z_n)\cdot K_{mm}\cdot(-z_n) - \lambda_n K_{nn} - (-z_n)\lambda_n z_n K_{mn} - (\xi-\lambda_n z_n) z_n K_{mn}$$
$$- (-z_n)V_m - z_n V_n + (-z_n z_m + 1)b$$

$$= (\xi z_n - \lambda_n)K_{mm} - \lambda_n K_{nn} + \lambda_n K_{mn} - (\xi z_n - \lambda_n)K_{mn}$$
$$+ z_n V_m - z_n V_n + (1 - z_n z_m)b$$

$$= -\lambda_n(K_{mm} + K_{nn} - 2K_{mn}) + \xi z_n K_{mm} - \xi z_n K_{mn}$$
$$+ z_n V_m - z_n V_n + (1 - z_n z_m)b$$

check $\frac{\partial^2 L_D(\lambda_n)}{\partial \lambda_n^2} = -(K_{mm} + K_{nn} - 2K_{mn})$ [suppose $\underline{u}_m = (a_1, \cdots, a_N)$, $\underline{u}_n = (b_1, \cdots, b_N)$]

$$= 2(a_1 b_1 + a_2 b_2 + \cdots + a_N b_N) - (a_1^2 + \cdots + a_N^2) - (b_1^2 + \cdots + b_N^2)$$

$$= -(a_1^2 - 2a_1 b + b_1^2) - (a_2^2 - 2a_2 b_2 + b_2^2) - \cdots - (a_N^2 - 2a_N b_N + b_N^2)$$

$$= -(a_1 - b_1)^2 - (a_2 - b_2)^2 - \cdots - (a_N - b_N)^2 \leq 0$$

$\therefore$ QED.

We already have

$$\frac{\partial L_D(\lambda_n)}{\partial \lambda_n} = -\lambda_n (K_{mm} + K_{nn} - 2K_{mn}) + \xi z_n K_{mm} - \xi z_n K_{mn} + z_n V_m - z_n V_n + (1 - z_n z_m) b$$

because $f(\underline{x}) = \underline{w}^T \underline{x} + w_0 = \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i^T \underline{x} + w_0 = \sum_{i=1}^{N} \lambda_i z_i K_{ix} + w_0$ ← prediction

$\underline{w}^* = \sum_{i=1}^{N} \lambda_i z_i \underline{u}_i$

Then
$$\begin{cases} V_m = \sum_{\substack{i=1 \\ (i \neq m,n)}}^{N} \lambda_i z_i K_{im} = f(\underline{x}_m) - \lambda_m z_m K_{mm} - \lambda_n z_n K_{nm} - w_0 \\ V_n = \sum_{\substack{i=1 \\ (i \neq m,n)}}^{N} \lambda_i z_i K_{in} = f(\underline{x}_n) - \lambda_m z_m K_{mn} - \lambda_n z_n K_{nn} - w_0 \end{cases}$$

$$\text{So} \quad V_m - V_n = f(\underline{x}_m) - f(\underline{x}_n) - \lambda_m z_m K_{mm} + \lambda_n z_n K_{nn} + (\lambda_m z_m - \lambda_n z_n) K_{mn}$$

$$\lambda_m = \varsigma z_m - \lambda_n z_n z_m$$

$$= f(\underline{x}_m) - f(\underline{x}_n) - (\varsigma z_m - \lambda_n z_n z_m) z_m K_{mm} + \lambda_n z_n K_{nn} + \left[(\varsigma z_m - \lambda_n z_n z_m) z_m - \lambda_n z_n \right] K_{mn}$$

$$= f(\underline{x}_m) - f(\underline{x}_n) - (\varsigma - \lambda_n z_n) K_{mm} + \lambda_n z_n K_{nn} + (\varsigma - 2\lambda_n z_n) K_{mn}$$

$$= \underbrace{f(\underline{x}_m) - f(\underline{x}_n)}_{\text{include } \lambda^{old}} - \varsigma K_{mm} + \varsigma K_{mn} + \underbrace{\lambda_n}_{\text{old } \lambda_n} z_n (K_{mm} + K_{nn} - 2K_{mn})$$

Now take it back to $\dfrac{\partial L_D(\lambda_n)}{\partial \lambda_n}$

$$\frac{\partial L_D(\lambda_n)}{\partial \lambda_n} = -\underbrace{\lambda_n}_{\text{new } \lambda_n} (K_{mm} + K_{nn} - 2K_{mn}) + \varsigma z_n K_{mm} - \varsigma z_n K_{mn} + z_n V_m - z_n V_n + (1 - z_n z_m) b = 0$$

$$= -\lambda_n^{new}(K_{mm}+K_{nn}-2K_{mn}) + \xi z_n(K_{mm}-K_{mn}) + (1-z_n z_m)b$$
$$+ z_n\left[f(\underline{x}_m) - f(\underline{x}_n) - \xi K_{mm} + \xi K_{mn} + \lambda_n^{old} z_n (K_{mm}+K_{nn}-2K_{mn})\right]$$

$$= -\lambda_n^{new}(K_{mm}+K_{nn}-2K_{mn}) + (1-z_n z_m)b + z_n\left[f(\underline{x}_m) - f(\underline{x}_n)\right]$$
$$+ \lambda_n^{old}(K_{mm}+K_{nn}-2K_{mn})$$

$$\frac{\partial L_D(\lambda_n)}{\partial \lambda_n} = 0 \quad \& \quad \text{let } \eta = K_{mm}+K_{nn}-2K_{mn} > 0 \text{ if } \underline{x}_m \neq \underline{x}_n$$

$$\Rightarrow \lambda_n^{new} = \lambda_n^{old} + \frac{(1-z_n z_m)b + z_n\left[f(\underline{x}_m) - f(\underline{x}_n)\right]}{\eta}$$

$$= \lambda_n^{old} + \frac{z_n\left[f(\underline{x}_m) - z_m b - f(\underline{x}_n) + z_n b\right]}{\eta} \qquad \text{Let } E_k = f(\underline{x}_k) - z_k b$$

Now $\boxed{\lambda_n^{new} = \lambda_n^{old} + \frac{z_n(E_m - E_n)}{\eta}}$

but we can't use λ_n^{new} to calculate λ_m^{new} for now.

Because λ_n^{new} may be out of the boundary $[0, \underset{\sim}{c}]$.

slack-SVM

Note: We cannot directly crop λ_n^{new} to $[0, c]$ because corresponding λ_m^{new} may still not satisfy $[0, c]$.

Ex: $\lambda_1 = 0.5$ $\quad \lambda_2 = 0.7$ $\quad \lambda_3 = 0.6$ $\quad \lambda_4 = 0.6$ $\quad C = 1$

$z_1 = 1$ $\quad z_2 = 1$ $\quad z_3 = -1$ $\quad \lambda_4 = -1$

Suppose we have new $\lambda_1 = -0.1$, cropped to 0

Then $0 + \lambda_2^{new} - 0.6 - 0.6 = 0 \implies \lambda_2^{new} = 1.2$ not satisfied!!!

v) Crop λ_n^{new}

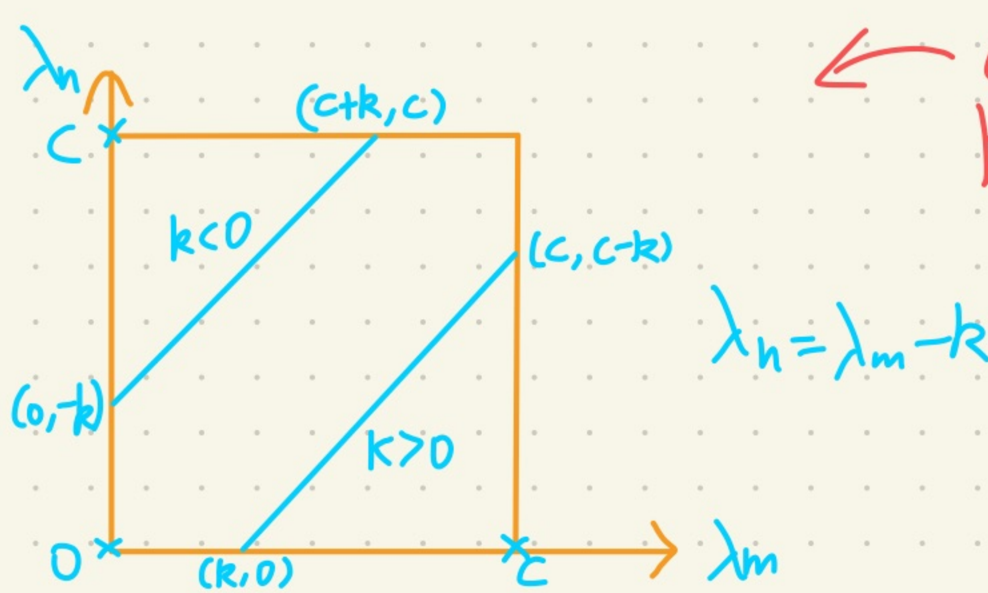
with $\begin{cases} \lambda_m z_m + \lambda_n z_n = \xi \\ 0 \le \lambda_m, \lambda_n \le C \end{cases}$ ← defined yourself [slack]

Case 1: if $z_m \neq z_n$

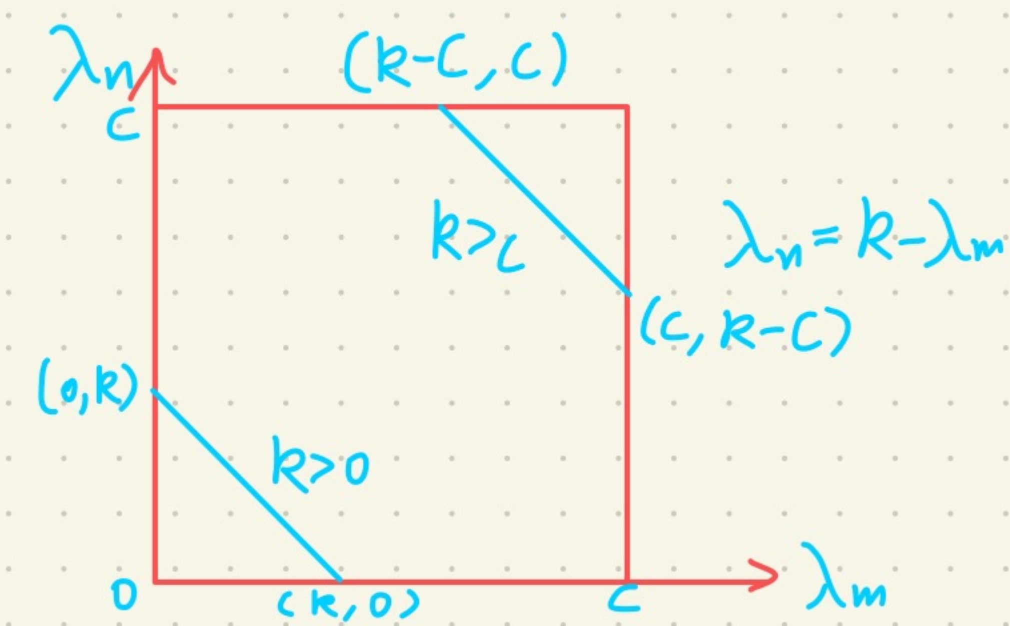
$$\lambda_m - \lambda_n = z_m \xi = k$$

Case 2: if $z_m = z_n$

$$\lambda_m + \lambda_n = z_m \xi = k$$



← calc new λ boundary by old λ's because old ones already satisfy the conditions



For λ_n:

$k<0 \quad -k \le \lambda_n \le C$

$k>0 \quad 0 \le \lambda_n \le C-k$

$\therefore$ we take

$\max(0,-k) \le \lambda_n \le \min(C, C-k)$

$$\underbrace{\max(0, \underbrace{\lambda_n-\lambda_m}_{old})}_{L_1} \le \lambda_n \le \underbrace{\min(C, C+\underbrace{\lambda_n-\lambda_m}_{old})}_{H_1}$$

For λ_n:

$k>0 \quad 0 \le \lambda_n \le k$

$k>C \quad k-C \le \lambda_n \le C$

$\therefore$ We take

$\max(0,k-C) \le \lambda_n \le \min(k, C)$

$$\underbrace{\max(0, \underbrace{\lambda_m+\lambda_n}_{old}-C)}_{L_2} \le \lambda_n \le \underbrace{\min(\underbrace{\lambda_m+\lambda_n}_{old}, C)}_{H_2}$$

Now we get the boundary where to fine-tune λ^{old} to λ^{new}.

$$[cropped]\lambda_n^{new} = \begin{cases} \text{if } z_m \ne z_n & \begin{cases} H_1 & \text{if } \lambda_n^{new} > H_1 \\ \lambda_n^{new} & \text{if } L_1 \le \lambda_n^{new} \le H_1 \\ L_1 & \text{if } \lambda_n^{new} < L_1 \end{cases} \\ \text{if } z_m = z_n & \begin{cases} \text{Similar, substitute } L_1, H_1 \text{ to } L_2, H_2 \end{cases} \end{cases}$$

vi) Calculate λ_m^{new}

Note: We don't need to use $\lambda_m^{new} = \xi z_m - \lambda_n^{new} z_n z_m$ to calculate λ_m^{new} because we have $\lambda_m^{old} = \xi z_m - \lambda_n^{old} z_n z_m$.

$$\Longrightarrow \boxed{\lambda_m^{new} = \lambda_m^{old} - z_n z_m (\lambda_n^{new} - \lambda_n^{old})}$$

vii) Calculate w_0^{new}

Reason: Because we use $w_0 \xrightarrow{calc} f(\underline{x}_k) \xrightarrow{calc} E_k \xrightarrow{calc} \lambda_n^{new}$, so we need to update w_0 for next iteration.

<u>Note</u>: We don't update $\underline{w}$ because $f(\underline{x}_k) = \sum_{i=1}^{N} \lambda_i z_i K_{ik} + w_0$, where

$\underline{w}$ had been substituted.

WHY don't we substitute w_0 before? Actually we can.

But $w_0^* = \frac{b}{z'} - \underline{w}^{*T} \cdot \underline{u}'$, where $\underline{u}'$ should be a support vector.

That means we need to find a support vector for old λ's, which is feasible but we just choose to calc w_0 later.

Now, we still use KKT to calculate w_0.

if $C > \lambda_m^{new}$ or $\lambda_n^{new} > 0$, we can use $\underline{x}_m$ or $\underline{x}_n$ to update w_0.

if λ_m^{new} and $\lambda_n^{new} = 0$ or C, $z_m(\underline{w}^T \underline{x}_m + w_0) \geq b$ make it $\stackrel{?}{=}$ equal zero $2w_0 = (z_m + z_n)b - \underline{w}^T \underline{x}_m - \underline{w}^T \underline{x}_n$

$z_n(\underline{w}^T \underline{x}_n + w_0) \geq b$ $\quad w_0 = \frac{(z_m+z_n)b - \underline{w}^T(\underline{x}_m + \underline{x}_n)}{2}$

So we can still use $\underline{x}_m, \underline{x}_n$ to satisfy KKT.

More specifically,

if $C > \lambda_m^{new} > 0$, $\quad z_m(\underline{w}^T \underline{x}_m + w_0) = b$

$$w_0^{new'} = b z_m - \underline{w}^T \underline{x}_m = b z_m - \sum_{i=1}^{N} \lambda_i z_i K_{im}$$

$$= b z_m - \sum_{\substack{i=1 \\ (i \neq m,n)}}^{N} \lambda_i z_i K_{im} - \lambda_m^{new} z_m K_{mm} - \lambda_n^{new} z_n K_{nm}$$

Because $E_m = f(\underline{x}_m) - z_m b = \underline{w}^T \underline{x}_m + w_0^{old} - z_m b$

$$= \sum_{\substack{i=1 \\ (i \neq m,n)}}^{N} \lambda_i z_i K_{im} + \lambda_m^{old} z_m K_{mm} + \lambda_n^{old} z_n K_{nm} + w_0^{old} - z_m b$$

Then $w_0^{new'} = \lambda_m^{old} z_m K_{mm} + \lambda_n^{old} z_n K_{nm} + w_0^{old} - \lambda_m^{new} z_m K_{mm} - \lambda_n^{new} z_n K_{nm} - E_m$

$$= w_0^{old} - E_m + z_m K_{mm}(\lambda_m^{old} - \lambda_m^{new}) + z_n K_{nm}(\lambda_n^{old} - \lambda_n^{new})$$

if $C > \lambda_n^{new} > 0$, similarly

$$w_0^{new2} = w_0^{old} - E_n + z_m K_{mn}(\lambda_m^{old} - \lambda_m^{new}) + z_n K_{nn}(\lambda_n^{old} - \lambda_n^{new})$$

if $C > \lambda_n^{new}, \lambda_m^{new} > 0$, $w_0^{new} = w_0^{new1} = w_0^{new2}$

if λ_n^{new} & $\lambda_m^{new} = 0$ or C,

we satisfy $\begin{cases} z_m(\underline{w}^T \underline{x}_m + w_0) \geq b \\ z_n(\underline{w}^T \underline{x}_n + w_0) \geq b \end{cases}$ by $\begin{cases} z_m(\underline{w}^T \underline{x}_m + w_0) = b \\ z_n(\underline{w}^T \underline{x}_n + w_0) = b \end{cases}$

$$\Longrightarrow 2w_0^{new} = bz_m - \underline{w}^T \underline{x}_m + bz_n - \underline{w}^T \underline{x}_n = w_0^{new1} + w_0^{new2}$$

$$\boxed{w_0^{new} = \frac{w_0^{new1} + w_0^{new2}}{2}}$$

Hence, whatever the condition is, we just calc $w_0^{new} = \dfrac{w_0^{new1} + w_0^{new2}}{2}$

Summary of Optimization Process

$E_k = f(\underline{x}_k) - z_k b$	$f(\underline{x}_k) = \sum_{i=1}^{N} \lambda_i^{old} z_i K_{ix} + w_0^{old}$
$\eta = K_{mm} + K_{nn} - 2K_{mn}$	$K_{pq} = \underbrace{\underline{u}_p^T \underline{u}_q}_{\text{linear Case}}$

$\underline{u}$ and $\underline{x}$ mean the same

For each iteration

1、Choose $m \neq n$ in $[1, N]$

2、Calculate $\lambda_n^{new} = \lambda_n^{old} + \dfrac{z_n(E_m - E_n)}{\eta}$

3、Crop λ_n^{new}

$$\begin{cases} \text{if } z_m \neq z_n & \max(0, \lambda_n^{old} - \lambda_m^{old}) \le \lambda_n^{new} \le \min(C, C + \lambda_n^{old} - \lambda_m^{old}) \\ \text{if } z_m = z_n & \max(0, \lambda_m^{old} + \lambda_n^{old} - C) \le \lambda_n^{new} \le \min(\lambda_m^{old} + \lambda_n^{old}, C) \end{cases}$$

4、Calculate $\lambda_m^{new} = \lambda_m^{old} - z_n z_m(\lambda_n^{new} - \lambda_n^{old})$

5、Calculate $w_0^{new} = \dfrac{w_0^{new1} + w_0^{new2}}{2}$

$$w_0^{new1} = w_0^{old} - E_m + z_m K_{mm}(\lambda_m^{old} - \lambda_m^{new}) + z_n K_{nm}(\lambda_n^{old} - \lambda_n^{new})$$

$$w_0^{new2} = w_0^{old} - E_n + z_m K_{mn}(\lambda_m^{old} - \lambda_m^{new}) + z_n K_{nn}(\lambda_n^{old} - \lambda_n^{new})$$

6、Next iteration

④ Kernel

$$(1+\underline{x}^T\underline{y})^2 = (1+x_0y_0+y_1y_1)^2 = 1+2x_0y_0+2x_1y_1+x_0^2y_0^2+2x_0x_1y_0y_1+x_1^2y_1^2$$
$$\simeq 1+x_0y_0+x_1y_1+x_0^2y_0^2+x_0x_1y_0y_1+x_1^2y_1^2$$
$$= [1, x_0, x_1, x_0^2, x_1^2, x_0x_1]\cdot[1, y_0, y_1, y_0^2, y_1^2, y_0y_1]^T$$

Ex:

- linear: $k(\underline{x}, \underline{x}') = \underline{x}^T\underline{x}'$
- quadratic: $(1+\underline{x}^T x')^{d=2} \xrightarrow{\text{simplify}} \underline{u} = [x_0, x_1, x_0x_1, x_0x_0, x_1x_1]$ if $\underline{x} = [x_0\ x_1]^T$
- Radial Basis Function (RBF): $\exp\{-\gamma\|\underline{x}-\underline{x}'\|_2^2\}$, $\gamma > 0$

Comment:

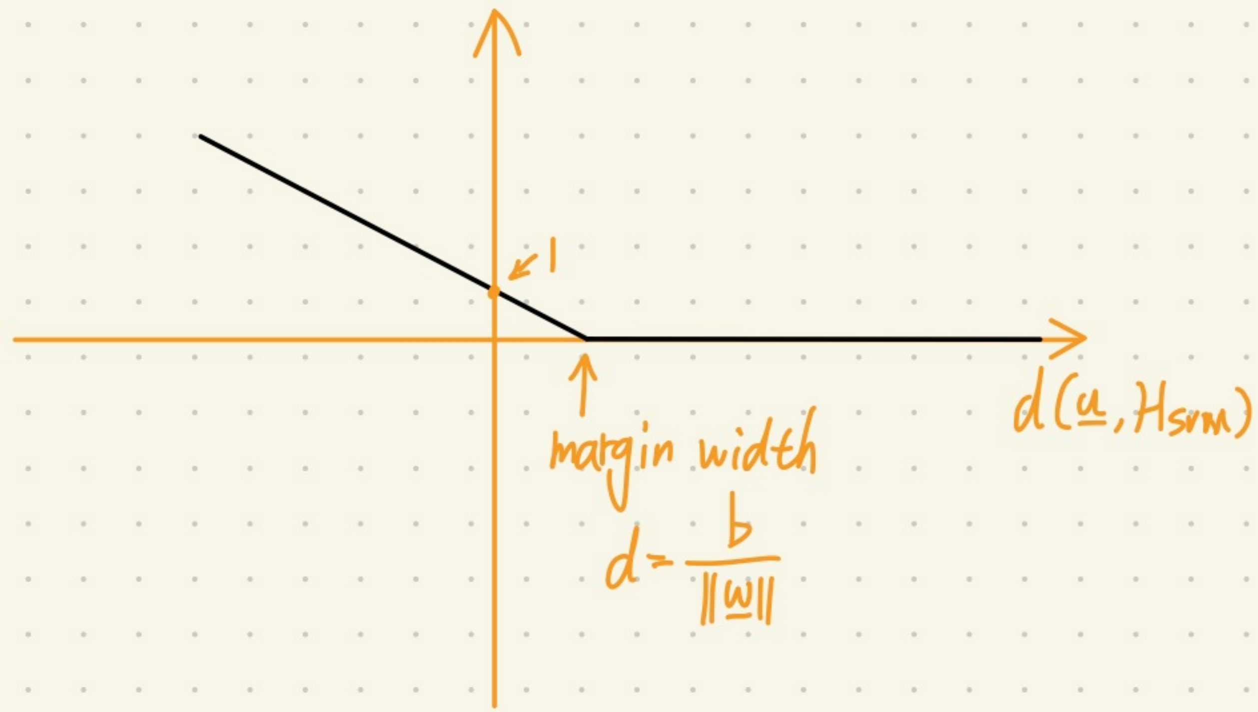
i) Often kernels are represented as $k(\underline{x}, \underline{x}') = \phi^T(\underline{x})\,\phi(\underline{x}')$.

For quadratic, it's easy to do it, so we can substitute $\underline{x}$ with $\underline{u} = \phi(\underline{x})$

For RBF, we have to substitute $\underline{x}^T\underline{x}'$ as a whole with $k(\underline{x}, \underline{x}')$.

Because for this mapping, $\phi(\underline{x})$ exists but has infinite dimensions!

5. slack variables ξ



① when $\underline{u}$ on correct side or margin line, $\xi = 0$

② when $\underline{u}$ within margin, $0 < \xi < 1$

③ when $\underline{u}$ on wrong side, $\xi > 1$

Hence, instead of $J = \frac{1}{2}\|\underline{w}\|$

we minimize

$$J = \frac{1}{2}\|\underline{w}\| + \underbrace{C}_{\text{slack parameter}} \cdot \sum_{i=1}^{N} \underbrace{\xi_i}_{\text{slack variable}}$$

Constraints: $z_i(\underline{w}^T \underline{u}_i + w_0) - b + \xi_i \geq 0$

Now, instead of $L(\underline{w}, w_0, \underline{\lambda})$

we use
$$L(\underline{w}, w_0, \underline{\xi}, \underline{\lambda}, \underline{\mu}) = \frac{1}{2}\|\underline{w}\|^2 + \underbrace{C\sum_{i=1}^{N}\xi_i}_{\text{new term}} - \sum_{i=1}^{N}\overset{\text{lagrange multiplier}}{\lambda_i}\left[z_i(\underline{w}^T\underline{u}_i + w_0) - b + \underbrace{\xi_i}_{\text{new term}}\right] - \underbrace{\sum_{i=1}^{N}\overset{\text{lagrange multiplier}}{\mu_i}\xi_i}_{\text{new term}}$$

$$\text{KKT}\begin{cases} \lambda_i \geq 0, \quad \lambda_i\left[z_i(\underline{w}^T\underline{u}_i + w_0) - b + \xi_i\right] = 0 & \forall i \\ z_i(\underline{w}^T\underline{u}_i + w_0) - b + \xi_i \geq 0 & \forall i \\ \underbrace{\mu_i \geq 0, \quad \mu_i\xi_i = 0, \quad \xi_i \geq 0}_{\text{new}} & \forall i \end{cases}$$

Similarly, calculate $\min\limits_{\underline{w}, w_0, \underline{\xi}} L(\underline{w}, w_0, \underline{\xi}, \underline{\lambda}, \underline{\mu})$

("max" for lagrange multipliers,
"min" for other parameters.)

$$\Rightarrow \begin{cases} \nabla_{\underline{w}} L = 0 \\ \nabla_{w_0} L = 0 \\ \nabla_{\underline{\xi}} L = 0 \end{cases}$$

$$\boxed{L = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\lambda_i\left[z_i(\underline{w}^T\underline{u}_i + w_0) - b + \xi_i\right] - \sum_{i=1}^{N}\mu_i\xi_i}$$

① $\nabla_{\underline{w}} L = 0$

$$\nabla_{\underline{w}} L(\underline{w}, w_0, \underline{\xi}, \underline{\lambda}, \underline{\mu}) = \underline{w}^* - \sum_{i=1}^{N}\lambda_i z_i \underline{u}_i = 0$$

$$\therefore \underline{w}^* = \sum_{i=1}^{N}\lambda_i z_i \underline{u}_i = 0$$

② $\nabla_{w_0} L = 0$

$$\nabla_{w_0} L = -\sum_{i=1}^{N}\lambda_i z_i = 0 \qquad \therefore \sum_{i=1}^{N}\lambda_i z_i = 0$$

③ $\nabla_{\xi_i} L = 0$

$$\nabla_{\xi_i} L = C - \lambda_i - \mu_i = 0 \qquad \therefore C = \lambda_i + \mu_i$$

take ① ② ③ back to $L(\underline{w}, w_0, \underline{\xi}, \underline{\lambda}, \underline{\mu})$

$$L_D(\underline{\lambda}, \underline{\mu}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T\underline{u}_j + \sum_{i=1}^{N}\lambda_i(b-\xi_i) - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T\underline{u}_j$$

$$- \underbrace{\sum_{i=1}^{N}\lambda_i z_i \cdot w_0}_{=0} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T\underline{u}_j + \sum_{i=1}^{N}\lambda_i b + \sum_{i=1}^{N}\underbrace{(-\lambda_i + C - \mu_i)}_{=0}\xi_i$$

$$L_D(\underline{\lambda}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T\underline{u}_j + \sum_{i=1}^{N}\lambda_i b$$

Obviously, the form is exactly the same as linear-separable case.

Comment: ① one of the difference between SVM & slack-SVM is the range of λ_i:

for SVM: $\lambda_i \geq 0$

for slack-SVM: $C \geq \lambda_i \geq 0$

because $C = \lambda_i + \mu_i$ and $\mu_i \geq 0$

so $\lambda_i \leq C$, which is determined by self.

② Another difference: w_0^{new}

In linearly-separable SVM, $\begin{cases} \text{if } \lambda_m^{new} \text{ or } \lambda_n^{new} \text{ in } (0, C),\ w_0^{new} = w_0^{new1} \text{ or } w_0^{new2} \\ \text{if neither, then } w_0^{new} = \dfrac{w_0^{new1} + w_0^{new2}}{2} \end{cases}$

In slack-SVM $\begin{cases} \text{if } \lambda_m^{new} \text{ or } \lambda_n^{new} \text{ in } (0, C), \text{ then } \xi_m \text{ or } \xi_n = 0,\ W_0^{new} = W_0^{new1} \text{ or } W_0^{new2} \\ \text{if neither, } \boxed{\text{we have to find one } \underline{u}_k \text{ to satisfy } \lambda_k \text{ in } (0, C)}. \end{cases}$

Because if not, we can't be sure where the point $\underline{u}_k$ is, and no idea what ξ_k is, then we can't use $z_k(\underline{w}^T \underline{x}_k + W_0) = b - \xi_k$ to get W_0^{new}

$\xi_k = 0$

$W_0 = b z_k - \underline{w}^T \underline{x}_k$

③ we don't need to know what $\underline{\xi}$ or $\underline{\mu}$ is.

④ C is slack-parameter and defined by us.

(or use cross validation to choose one)

C is larger, false tolerance is fewer.